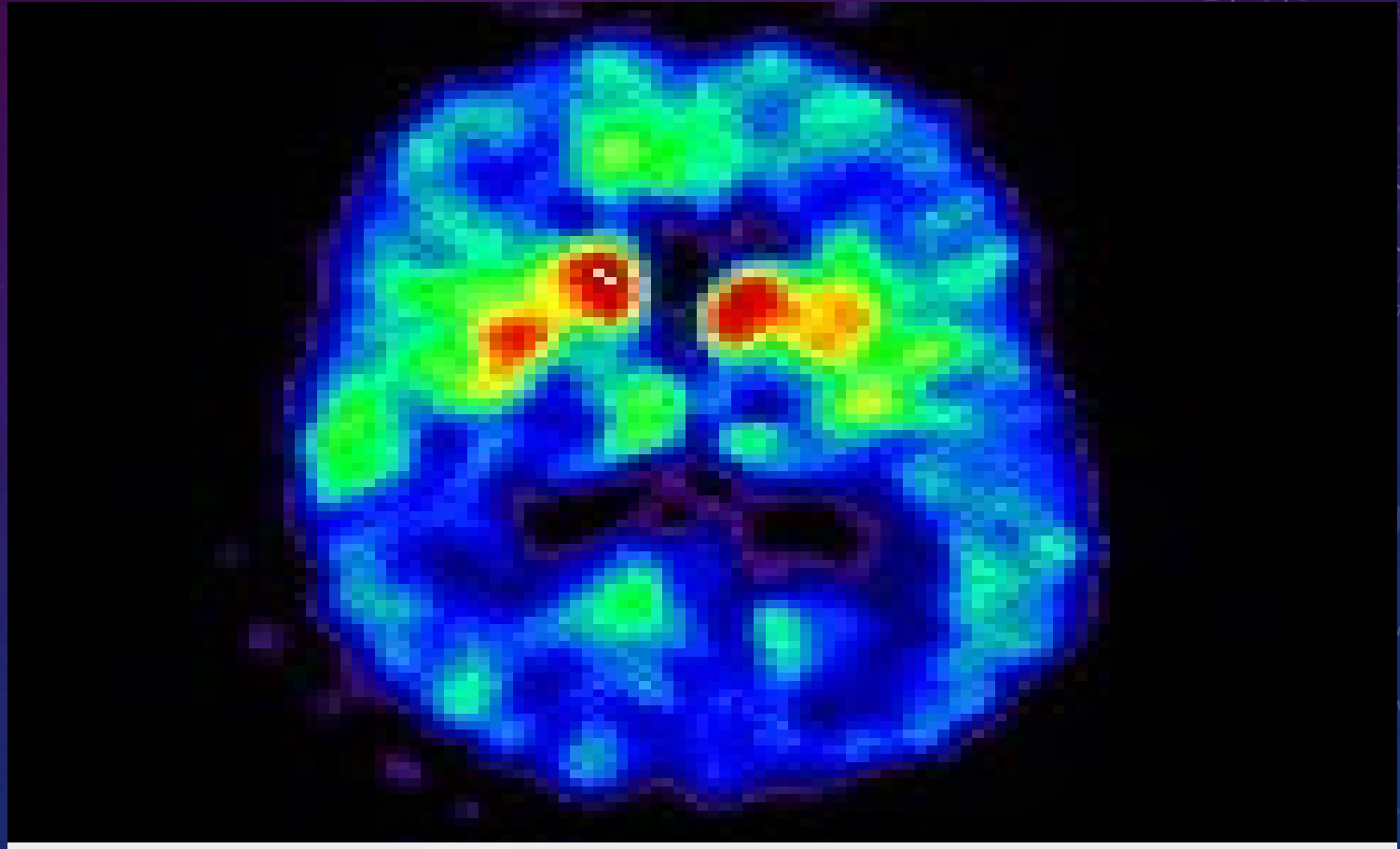


PROJET DE VISUALISATION DE DONNÉES



- SERIK MOHAMED : 11707386

- ANNÉE UNIVERSITAIRE:2019/2020

DONNÉES MALADIE DE PARKINSON

Sommaire

1-Introduction

2-exploration et visualisation avec KNIME

- Les données d'un état initial (outil statistics)

- L'outil Linear Correlation

- L'outil Correlation filter

- L'outil PCA

- L'outil MDS

- L'outil k-means

- L'outil k-means+PCA

- Hierarchical clustering

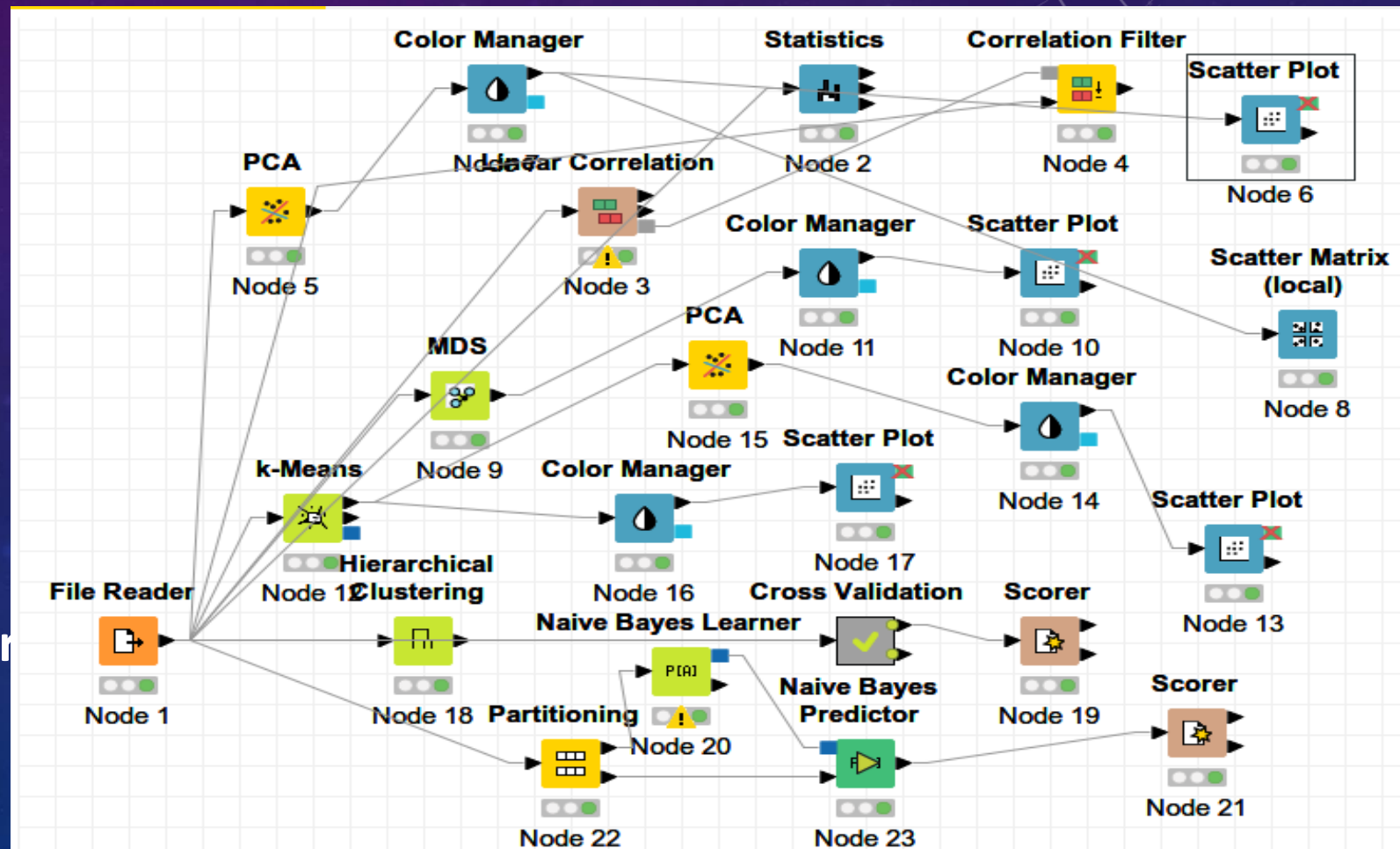
- L'outil Cross validation

- l'outil naive bayes learner

- l'outil naive bayes predictor

- l'outil scatter matrix

3-conclusion

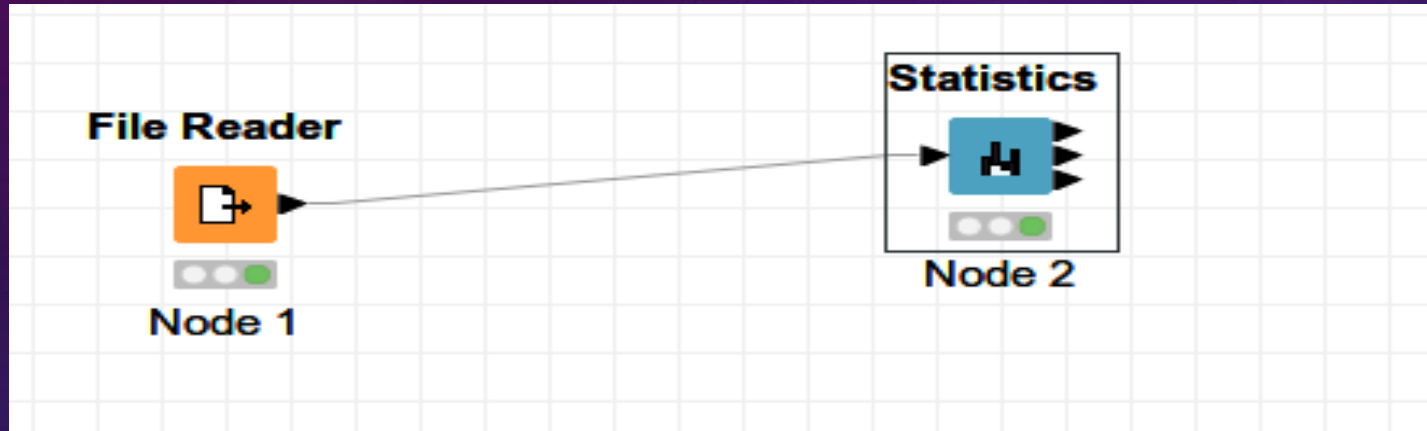


1-Introduction

Cette base de données est composée d'une série de mesures biomédicales de la voix à partir de 31 personnes, 23 ayant la maladie de Parkinson (MP). Chaque colonne de la table est une mesure particulière de la voix, et chaque ligne correspond à une des 195 enregistrements de ces personnes (colonne "nom"). L'objectif principal des données est de discriminer les personnes en bonne santé de ceux avec MP, selon "l'état" colonne qui est fixé à 0 pour la santé et une pour MP.

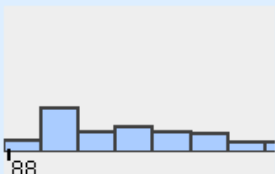
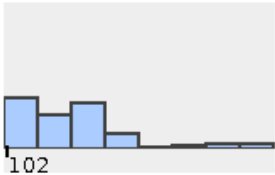
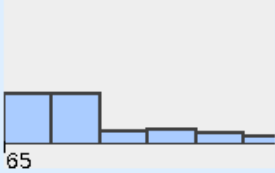
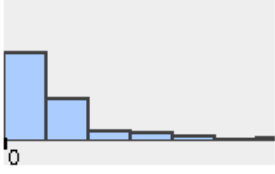
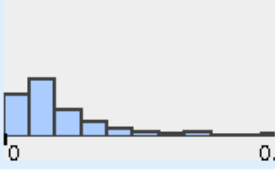
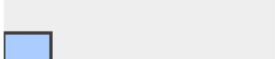
2-exploration et visualisation avec KNIME

-Les données d'un état initial (outil statistics)

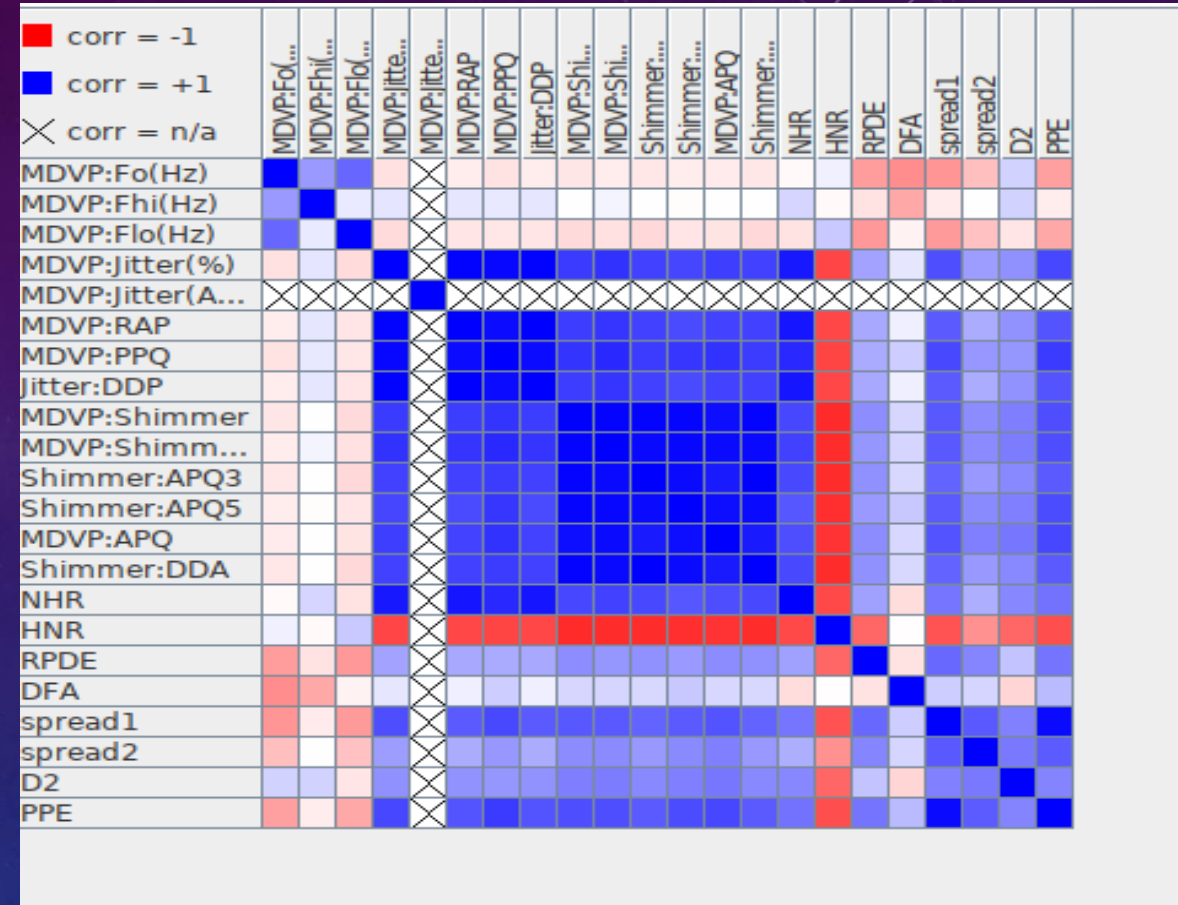
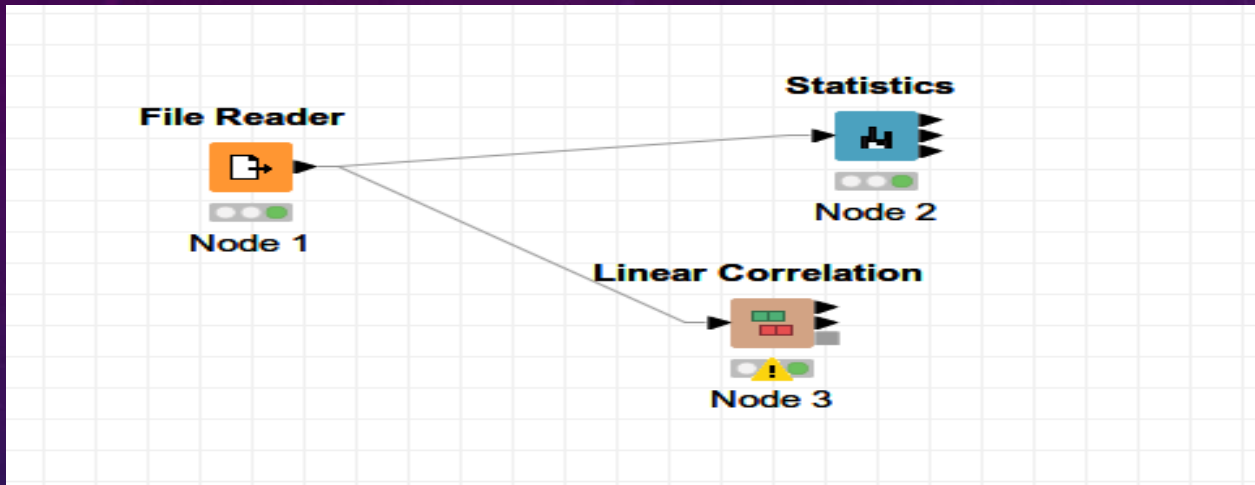


Ce nœud calcule des moments statistiques tels que minimum, maximum, moyenne, écart-type, variance, médiane, somme globale, nombre de valeurs manquantes et nombre de lignes dans toutes les colonnes numériques, et compte toutes les valeurs nominales ainsi que leurs Occurrences.

Voici le résultat d'applications de l'outil statistics

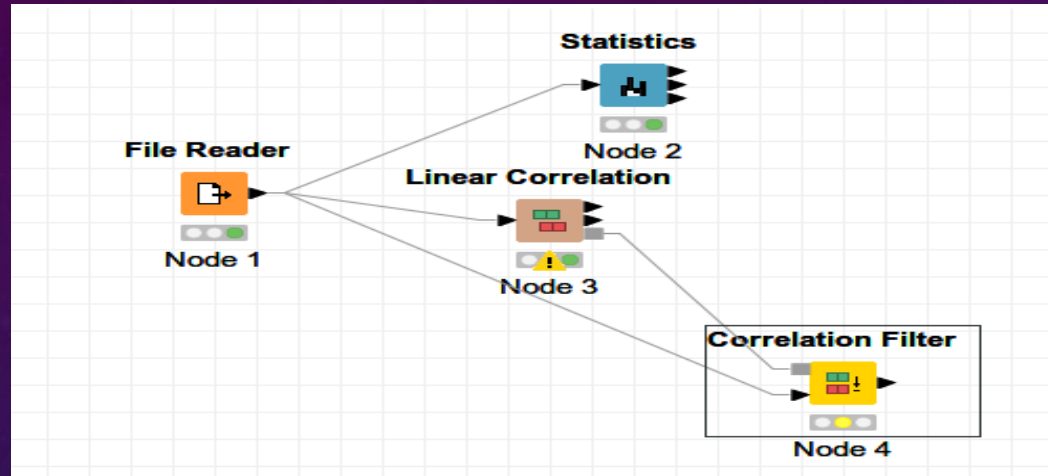
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
MDVP:Fo(Hz)	88,333	154,2286	148,79	260,105	41,3901	0,5917	-0,6279	0	0	0	
MDVP:Fhi(Hz)	102,145	197,1049	175,829	592,03	91,4915	2,5421	7,6272	0	0	0	
MDVP:Flo(Hz)	65,476	116,3246	104,315	239,17	43,5214	1,2174	0,6546	0	0	0	
MDVP:Jitter(%)	0,0017	0,0062	0,0049	0,0332	0,0048	3,0849	12,0309	0	0	0	
MDVP:Jitter (Abs)	7,00E-6	4,40E-5	3,00E-5	0,0003	3,48E-5	2,6491	10,869	0	0	0	
MDVP:RAP	0,0007	0,0033	0,0025	0,0214	0,003	3,3607	14,2138	0	0	0	

-L'outil Linear Correlation

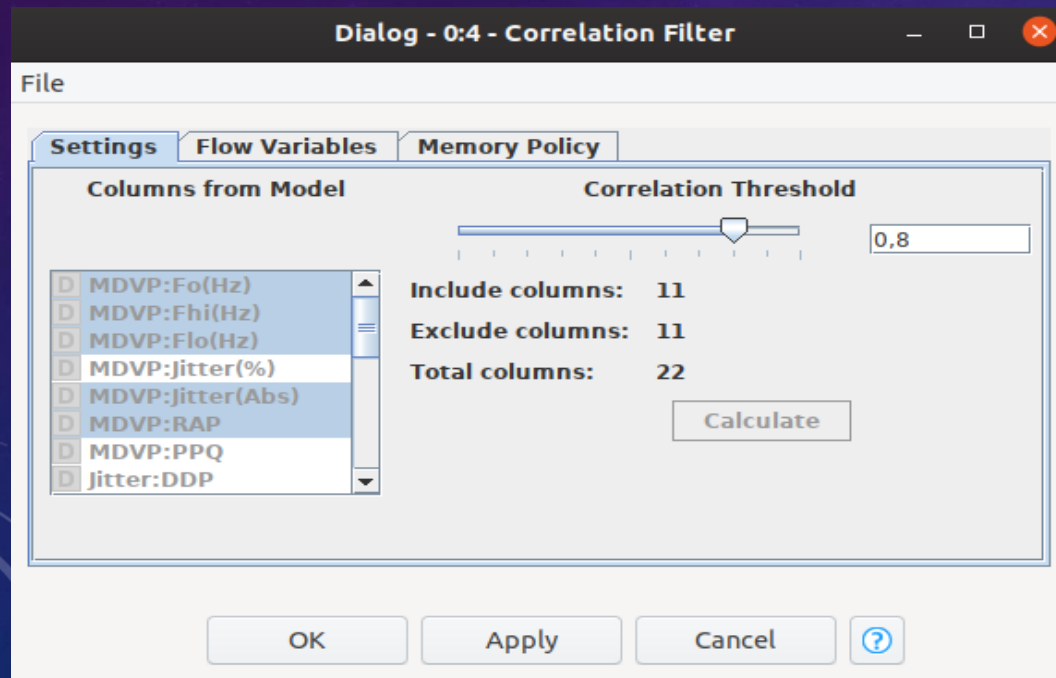


Ça nous permet d'obtenir cette matrice de corrélation sur laquelle on a constaté trop de corrélation entre les données à cause des mêmes informations communes qui existe entre les personnes.

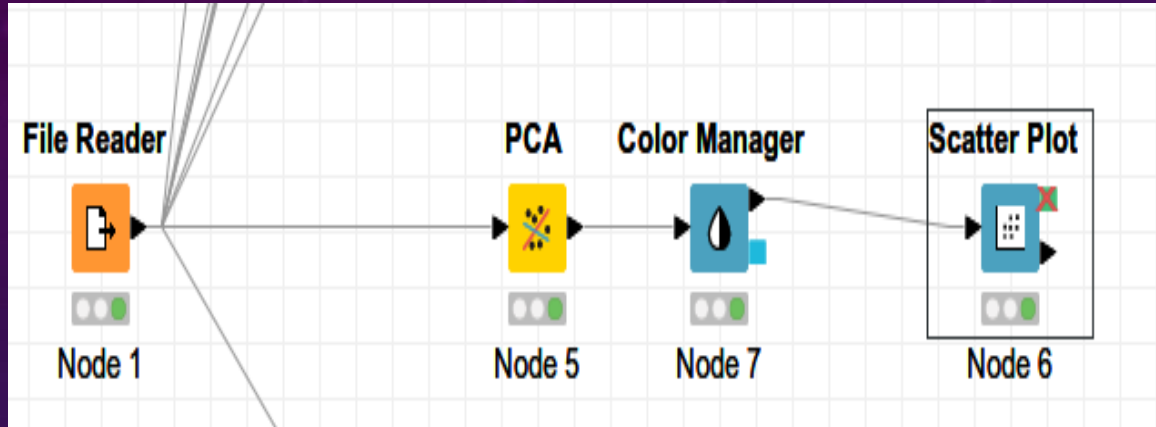
-L'outil Correlation Filter



Le nœud Filtre de corrélation recherche la colonne avec les colonnes les plus corrélées dans une procédure itérative. Le tableau de sortie contiendra l'ensemble de colonnes réduit en fonction de seuil choisi (0,8 dans notre cas).



-L'outil PCA



Consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées «composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

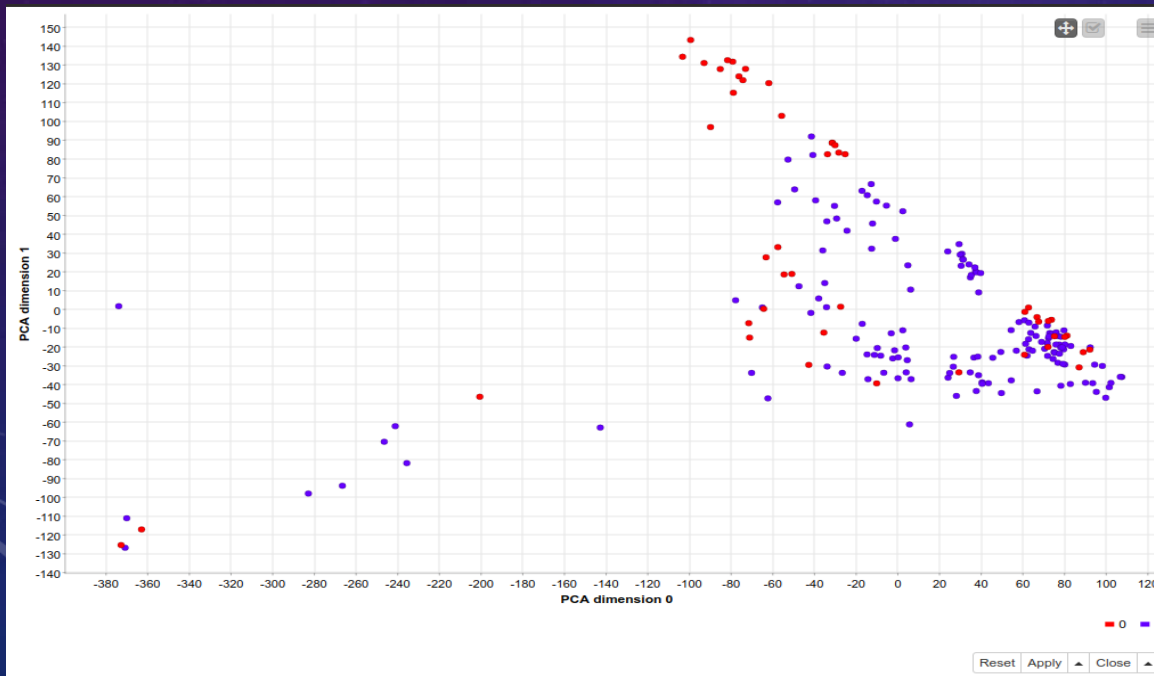


Tableau de dialogue PCA à 2 dimensions

Dialog - 0:5 - PCA

File

Settings Flow Variables Memory Policy

Target dimensions

☒ Dimension(s) to reduce to

☐ Minimum information fraction

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

No columns in this list

☒ Enforce exclusion

Include

- D MDVP:F0(Hz)
- D MDVP:F1(Hz)
- D MDVP:F0(Hz)
- D MDVP:Jitter(%)
- D MDVP:Jitter(Abs)
- D MDVP:RAP
- D MDVP:PPQ
- D Jitter:DDP
- D MDVP:Shimmer

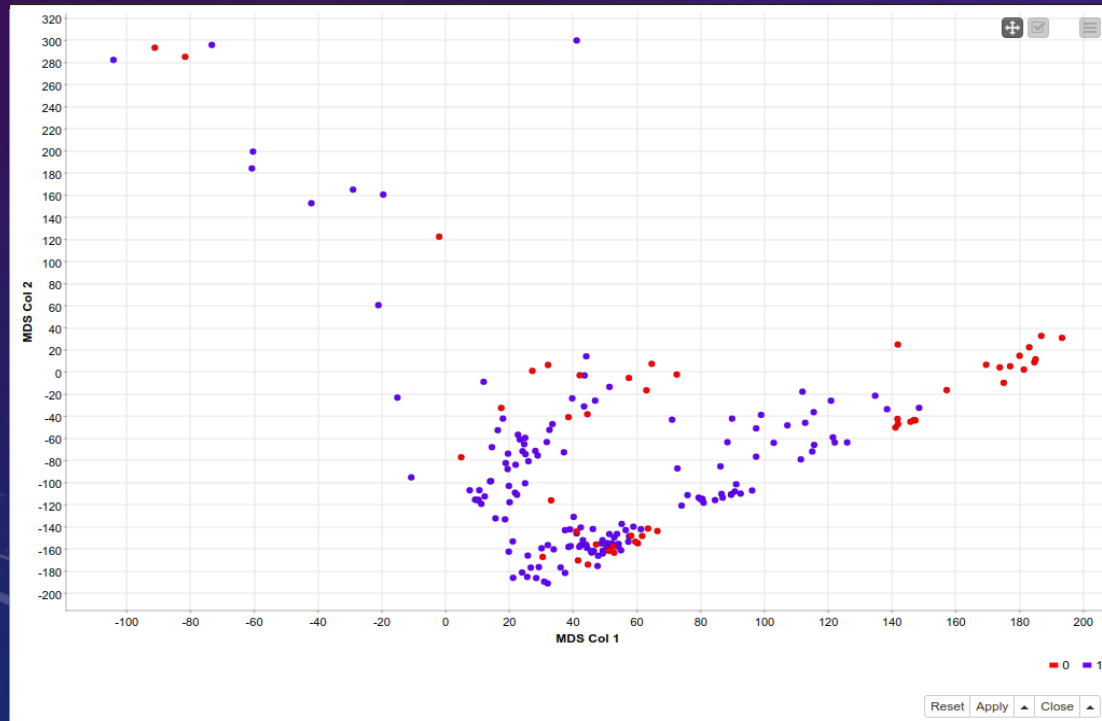
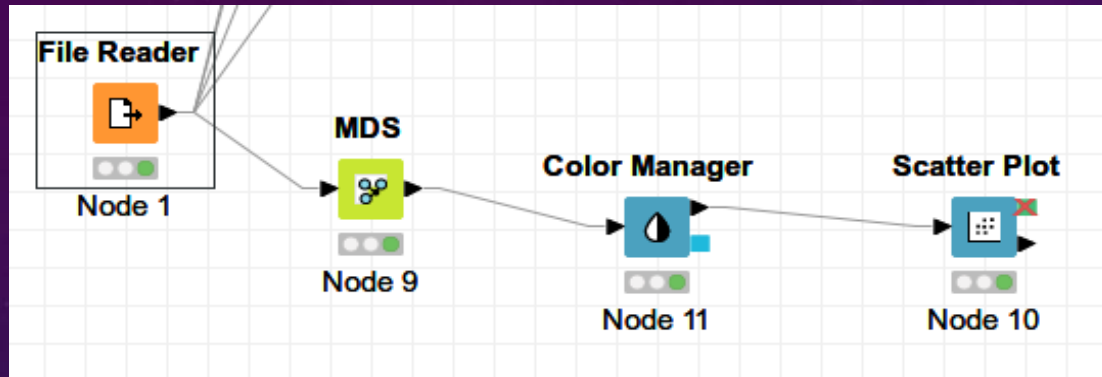
☐ Enforce inclusion

☐ Remove original data columns

☐ Fail if missing values are encountered

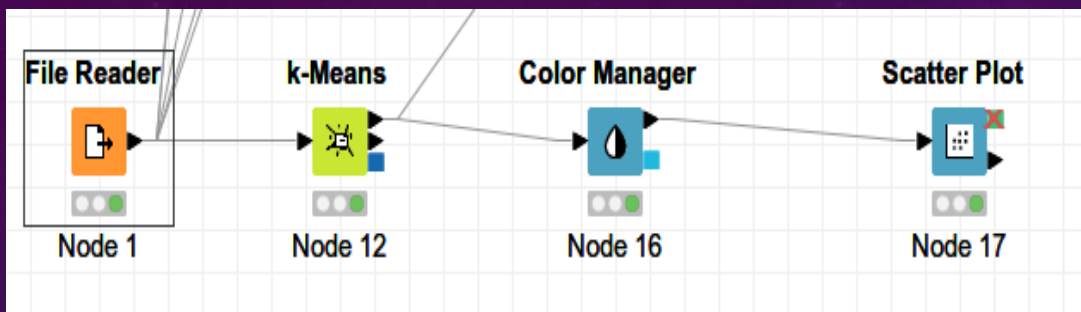
OK Apply Cancel ?

-L'outil MDS « Multidimensional scaling »

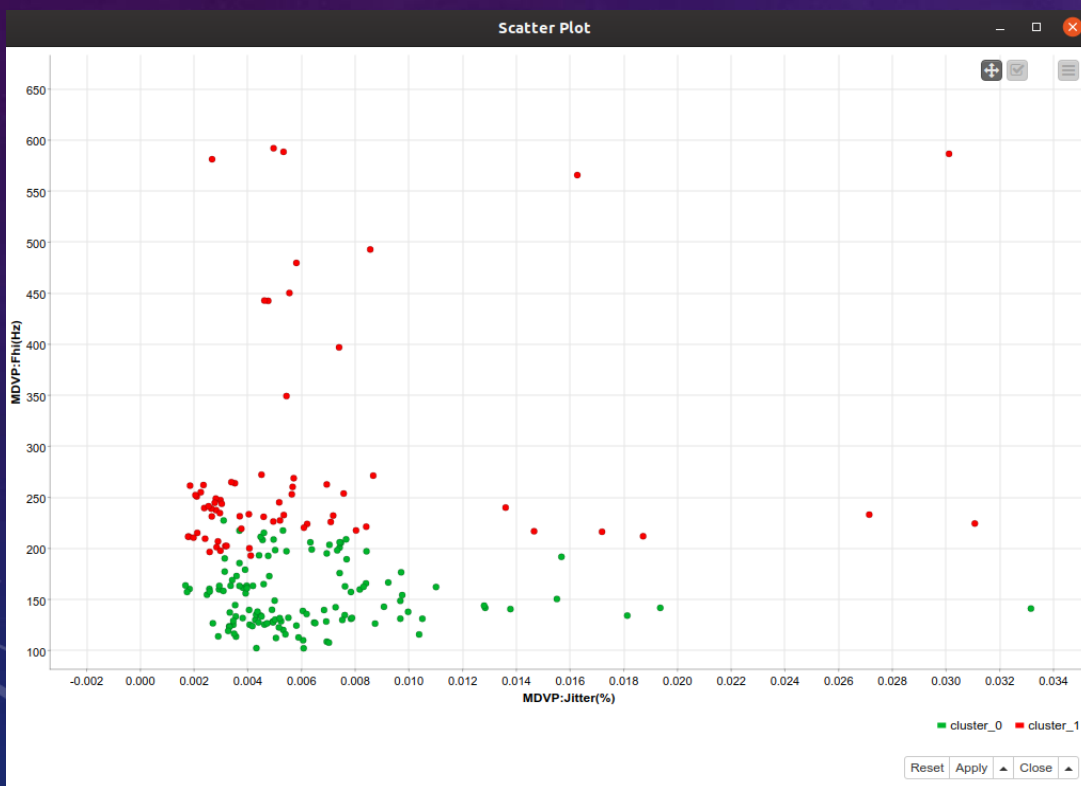


C'est un ensemble de techniques statistiques utilisées pour explorer les similarités dans les données. Étant donné N points X dans un espace de dimension p , le positionnement multidimensionnel consiste à représenter ces points dans un espace de dimension $m < p$ par N nouveaux points Y en conservant les proximités. Présenté dans ce contexte, le positionnement multidimensionnel est une technique de réduction de dimension, au même titre que PCA.

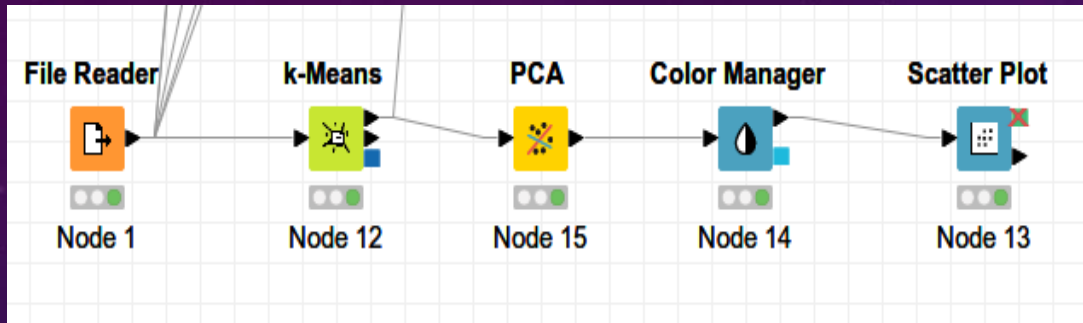
L'outil K-means



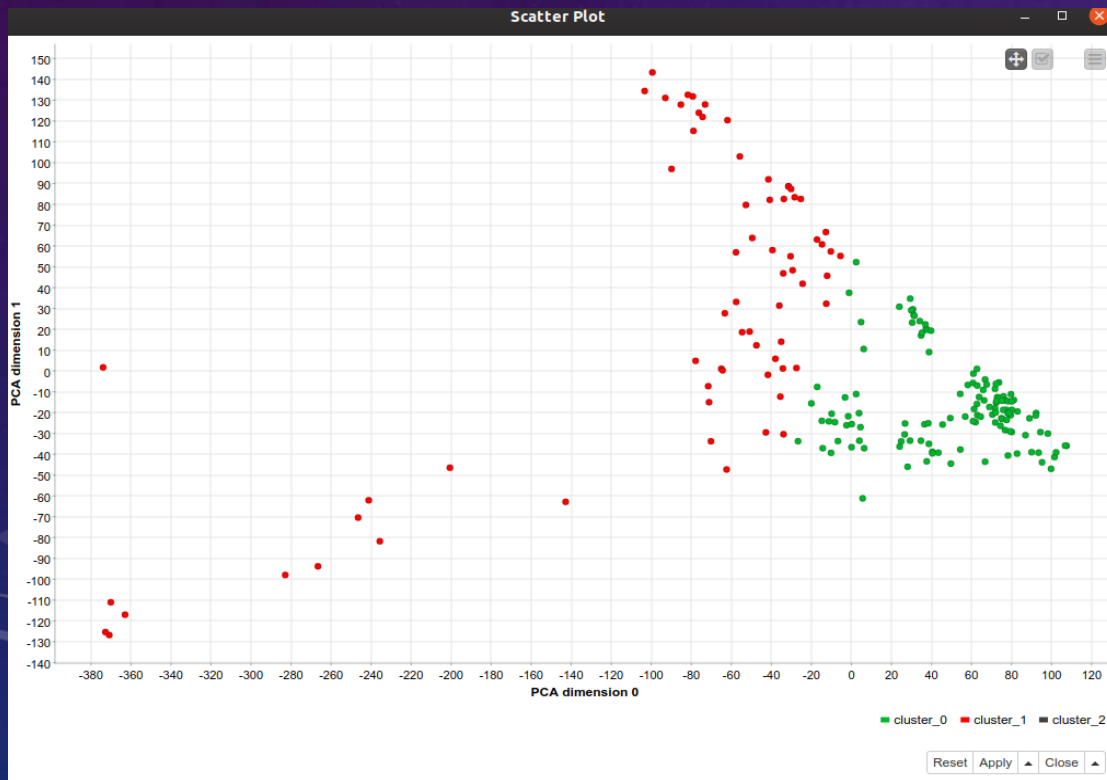
Le clustering k -means est une méthode de quantification vectorielle issue du traitement du signal , très utilisées pour l'analyse de cluster dans l'exploration de données. Elle consiste à grouper un ensemble d'objets de telle sorte que les objets du même groupe (appelé cluster) se ressemblent davantage que ceux d'autres groupes (clusters).



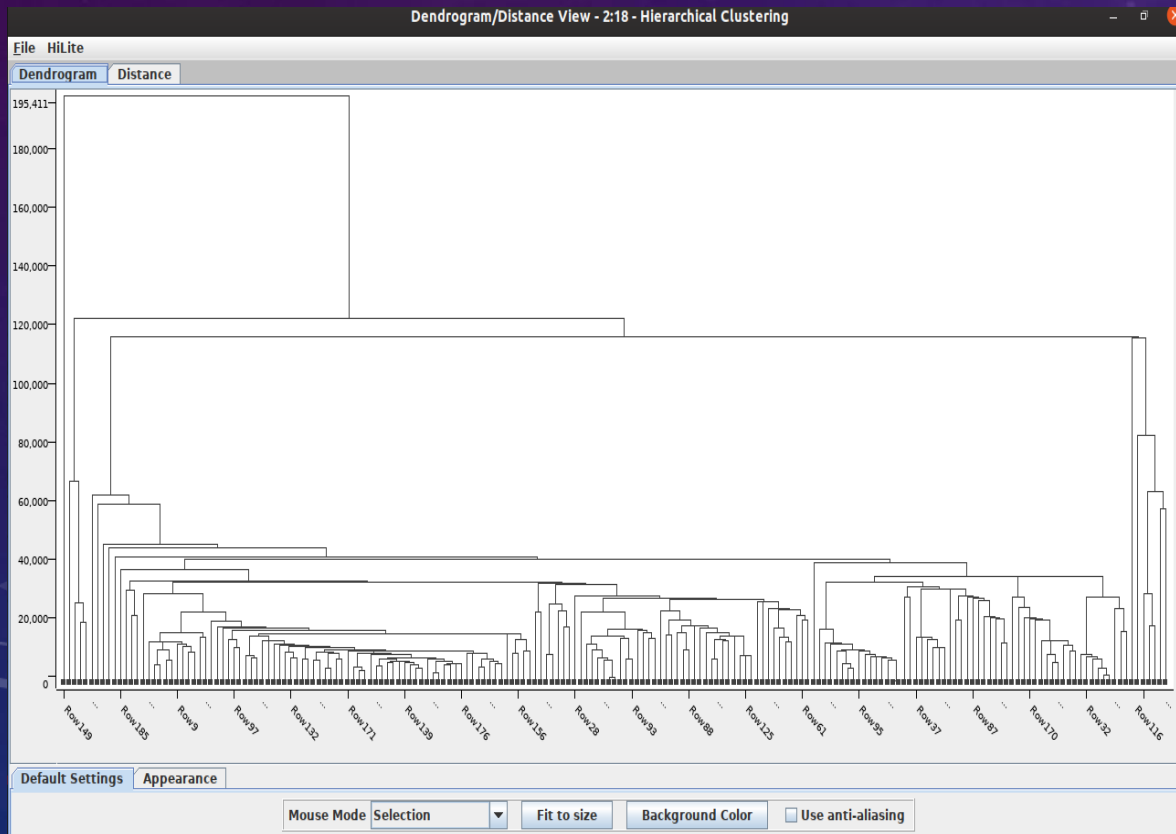
L'outil k-means+PCA



L'outil k-means+PCA permet de mieux regrouper les données et les repartir

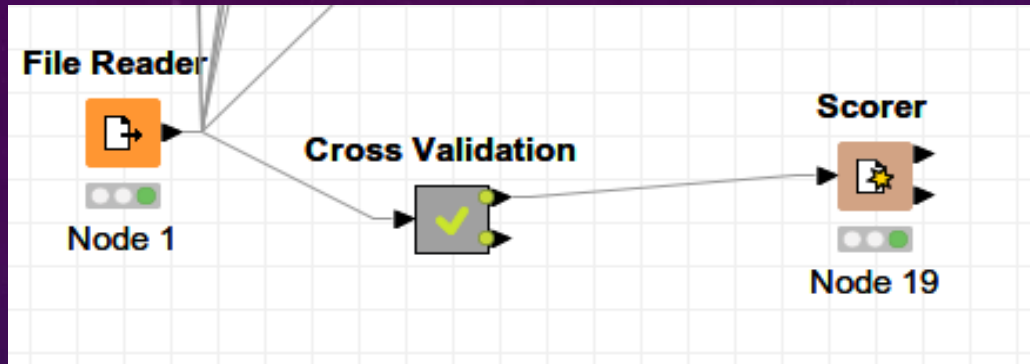


-Hierarchical clustering



On cherche à réduire le nombre de classes n à $\text{nb_classes} < n$, ceci se fait itérativement. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus « proches », en d'autres termes, celles dont la dissimilarité entre elles est minimale, cette valeur de dissimilarité est appelée indice d'agrégation. Comme on rassemble d'abord les individus les plus proches, la première itération a un indice d'agrégation faible, mais celui-ci va croître d'itération en itération.

L'outil cross validation



Fournit un squelette de nœuds nécessaires à la validation nœuds nécessaires à la validation croisée

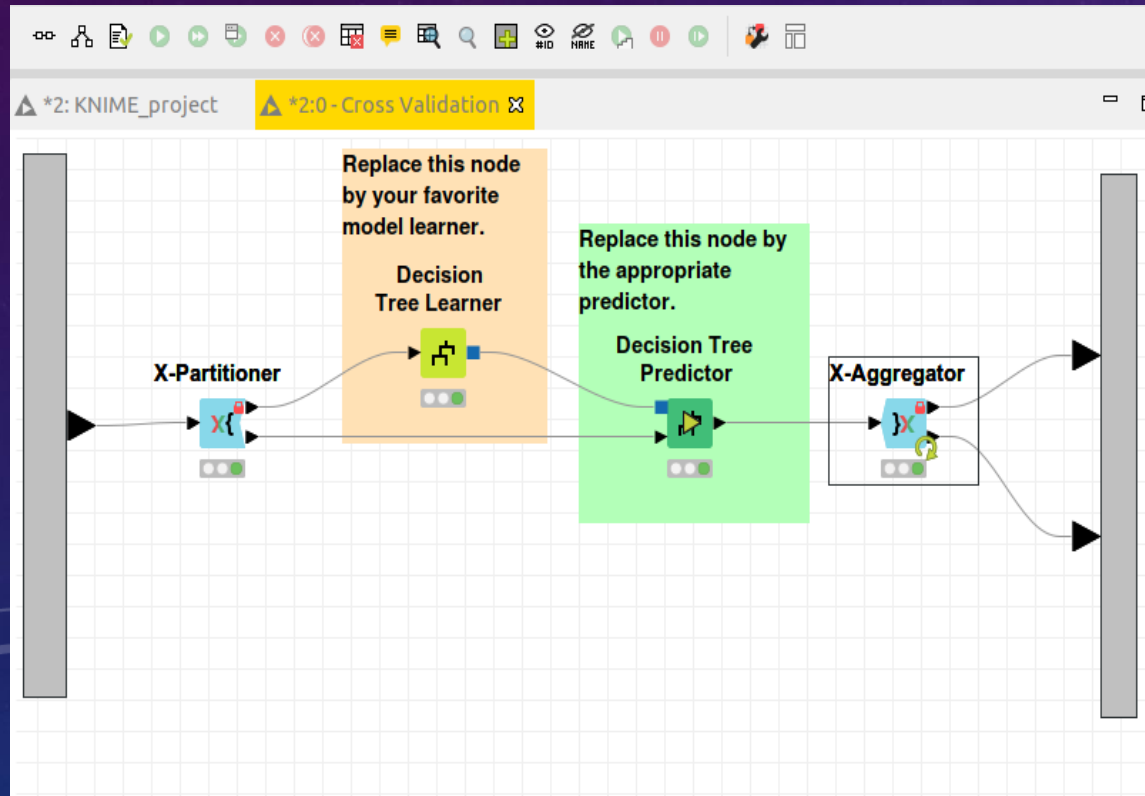
Nœuds contenus:

X-Aggregator: Nœud qui agrège le résultat pour la validation croisée.

X-Partitioner: Partitionneur de données à utiliser dans un flux de validation

Decision Tree Predictor: Utilise un arbre de décision existant pour calculer les étiquettes de classe pour les vecteurs d'entrée.

Decision Tree Learner: Induction d'arbre de décision effectuée en mémoire.



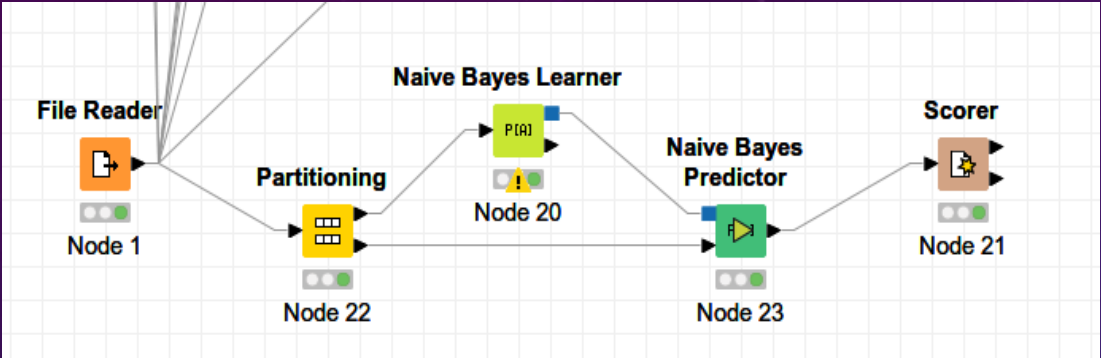
Confusion Matrix - 2:19 - Scorer

File Hilite	1	0
status \ Pr...		
1	132	15
0	12	36

Correct classified: 168
Accuracy: 86,154 %
Wrong classified: 27
Error: 13,846 %

-L'outil Naive Bayes learner

Le nœud crée un modèle bayésien à partir des données d'apprentissage fournies. Il calcule le nombre de lignes par valeur d'attribut par classe pour les attributs nominaux et la distribution gaussienne pour les attributs numériques.



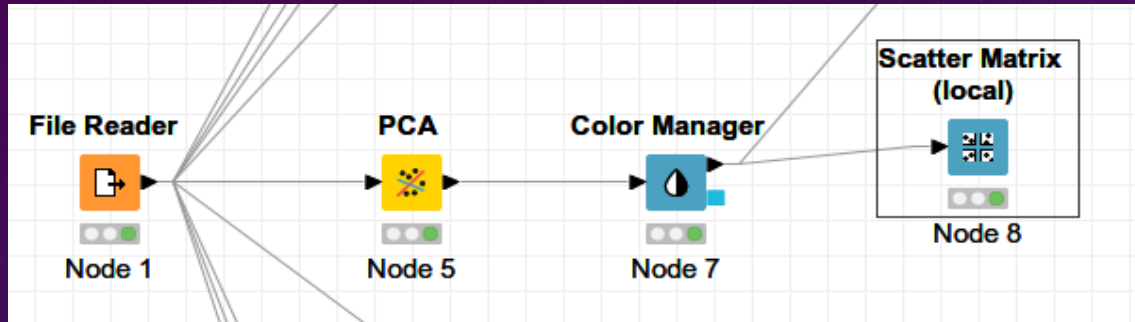
File		
⚠ The following attributes are skipped: name/Too many values		
Class counts for status		
Class:	0	1
Count:	33	103
Total count: 136		
Threshold to used for zero probabilities: 1.0E-4		
Skipped attributes: name/Too many values		
Gaussian distribution for D2 per class value		
	0	1
Count:	33	103
Mean:	2.1567	2.47187
Std. Deviation:	0.29888	0.37722
Rate:	24 %	76 %
Gaussian distribution for DFA per class value		
	0	1
Count:	33	103
Mean:	0.69577	0.71889
Std. Deviation:	0.05238	0.05339
Rate:	24 %	76 %
Gaussian distribution for HNR per class value		
	0	1

-L'outil Naive Bayes predictor

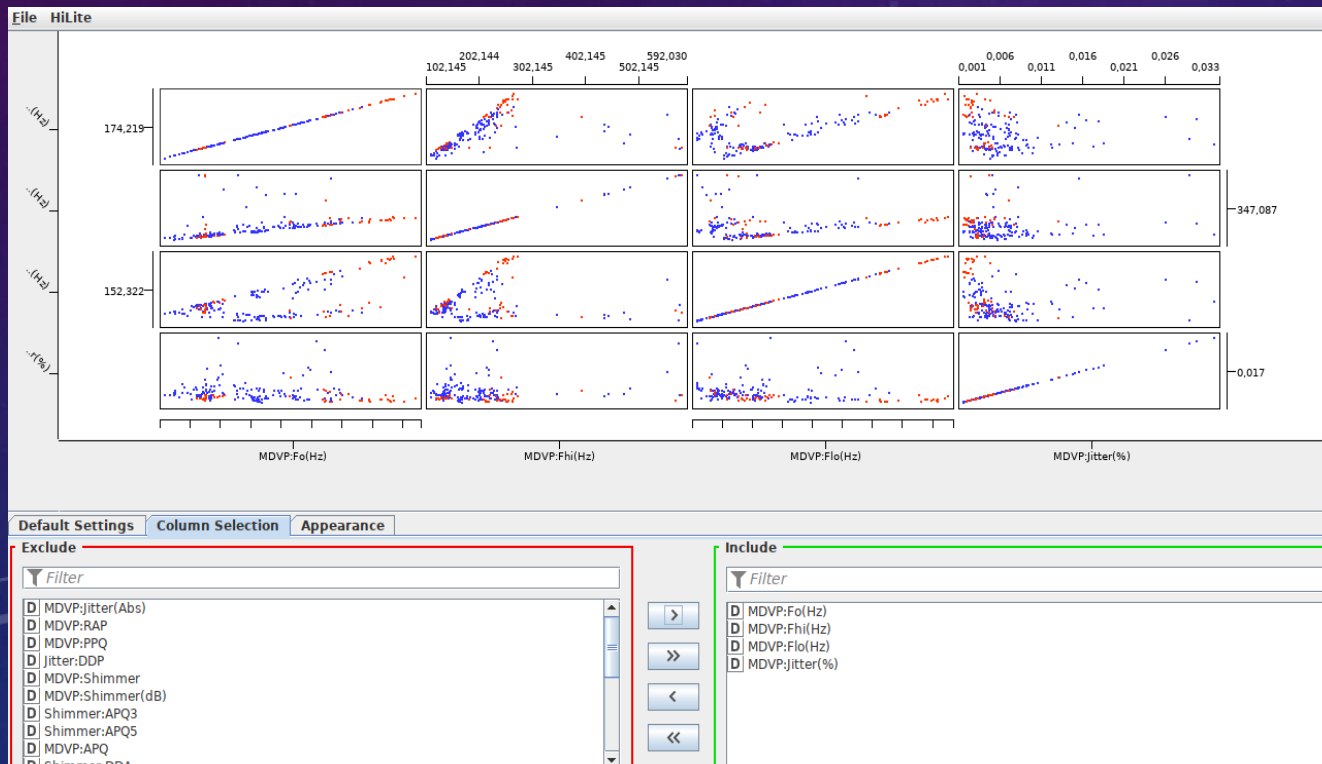
File Hilite Navigation View																		
Table "default" - Rows: 59			Spec - Columns: 25		Properties		Flow Variables											
Row ID	MDVP...	D MDVP...	D Shim...	D Shim...	D MDVP...	D Shim...	D NHR	D HNR	S status	D RPDE	D DFA	D spread1	D spread2	D D2	D PPE	S Predic...		
Row0	4	0.426	0.022	0.031	0.03	0.065	0.022	21.033	1	0.415	0.815	-4.813	0.266	2.301	0.285	1		
Row1	1	0.626	0.031	0.045	0.044	0.094	0.019	19.085	1	0.458	0.82	-4.075	0.336	2.487	0.369	1		
Row4	4	0.584	0.035	0.048	0.045	0.105	0.018	19.649	1	0.417	0.823	-3.748	0.235	2.332	0.41	1		
Row7	6	0.134	0.008	0.009	0.013	0.025	0.003	26.892	1	0.637	0.763	-6.168	0.184	2.065	0.164	0		
Row8	1	0.191	0.011	0.013	0.017	0.032	0.011	21.812	1	0.616	0.774	-5.499	0.328	2.323	0.232	1		
Row9	8	0.255	0.014	0.017	0.024	0.043	0.01	21.862	1	0.547	0.798	-5.012	0.326	2.433	0.271	1		
Row19		0.348	0.017	0.024	0.043	0.052	0.034	17.153	1	0.65	0.686	-4.554	0.34	2.857	0.322	1		
Row24	7	0.164	0.007	0.009	0.015	0.022	0.018	23.831	1	0.398	0.732	-5.557	0.221	2.692	0.216	0		
Row33		0.085	0.005	0.006	0.007	0.014	0.001	32.684	0	0.369	0.742	-7.696	0.179	1.545	0.056	0		
Row34		0.085	0.005	0.006	0.007	0.014	0.001	33.047	0	0.34	0.742	-7.965	0.164	1.423	0.045	0		
Row38	4	0.131	0.007	0.009	0.012	0.022	0.002	26.738	1	0.404	0.766	-6.452	0.212	2.269	0.142	0		
Row42	8	0.164	0.01	0.01	0.011	0.031	0.007	22.736	0	0.305	0.654	-7.311	0.099	2.417	0.095	0		
Row43	8	0.154	0.01	0.01	0.013	0.03	0.007	23.145	0	0.458	0.634	-6.794	0.158	2.257	0.117	0		
Row51		0.185	0.011	0.01	0.016	0.034	0.005	25.03	0	0.508	0.76	-6.689	0.292	2.432	0.106	0		
Row53	5	0.228	0.014	0.012	0.019	0.042	0.005	25.429	0	0.42	0.786	-6.837	0.27	2.224	0.147	0		
Row54	7	0.255	0.015	0.016	0.021	0.044	0.01	21.028	1	0.536	0.819	-4.65	0.206	1.987	0.317	1		
Row56	7	0.334	0.02	0.023	0.028	0.061	0.01	21.422	1	0.542	0.821	-4.438	0.238	1.923	0.335	1		
Row57	3	0.221	0.012	0.015	0.019	0.036	0.007	22.817	1	0.531	0.818	-4.608	0.29	2.022	0.314	1		
Row59	2	0.35	0.018	0.02	0.025	0.054	0.011	21.66	1	0.548	0.817	-4.609	0.222	1.832	0.316	1		
Row60	9	0.17	0.01	0.013	0.014	0.029	0.009	25.554	0	0.342	0.679	-7.041	0.067	2.461	0.102	0		
Row63	6	0.145	0.009	0.011	0.012	0.026	0.004	25.964	0	0.257	0.683	-7.246	0.019	2.498	0.094	0		
Row65	7	0.154	0.009	0.012	0.012	0.028	0.004	24.547	0	0.305	0.682	-7.314	0.006	2.119	0.092	0		
Row69	4	0.497	0.034	0.025	0.036	0.101	0.024	21.718	1	0.487	0.727	-6.261	0.121	2.137	0.142	1		
Row75	2	0.206	0.013	0.012	0.017	0.039	0.005	25.197	1	0.464	0.807	-5.478	0.315	1.862	0.229	0		
Row81	2	0.296	0.018	0.018	0.022	0.054	0.018	19.659	1	0.576	0.779	-5.132	0.21	2.233	0.261	1		
Row82	4	0.216	0.014	0.014	0.018	0.041	0.012	20.536	1	0.555	0.788	-5.022	0.147	2.428	0.265	1		
Row83	2	0.202	0.013	0.013	0.016	0.038	0.009	22.244	1	0.577	0.772	-6.025	0.078	2.054	0.177	0		
Row85	9	0.331	0.021	0.025	0.029	0.063	0.028	16.176	1	0.584	0.728	-5.658	0.316	3.098	0.2	1		
Row87	5	0.58	0.037	0.042	0.047	0.11	0.048	13.922	1	0.603	0.741	-5.515	0.3	3.137	0.221	1		
Row93	9	0.637	0.033	0.038	0.044	0.1	0.012	20.969	1	0.447	0.698	-6.153	0.174	2.08	0.161	1		
Row94	7	0.307	0.021	0.023	0.028	0.062	0.009	22.219	1	0.502	0.712	-6.251	0.188	2.144	0.161	1		
Row95	3	0.283	0.018	0.019	0.026	0.054	0.01	21.693	1	0.447	0.706	-6.247	0.181	2.344	0.165	1		
Row97	9	0.342	0.021	0.022	0.031	0.064	0.082	15.338	1	0.63	0.714	-4.02	0.265	2.672	0.341	1		
Row99	7	0.659	0.036	0.04	0.058	0.108	0.167	12.435	1	0.639	0.675	-3.76	0.243	2.635	0.378	1		
Row108		0.093	0.005	0.006	0.01	0.014	0.002	29.928	1	0.311	0.676	-6.739	0.161	2.297	0.115	0		
Row114	1	0.209	0.011	0.013	0.016	0.033	0.01	21.864	1	0.332	0.715	-6.73	0.182	2.938	0.116	0		

Prédit la classe par ligne en fonction du modèle appris. La probabilité de classe est le produit de la probabilité par attribut et de la probabilité de l'attribut de classe lui-même.

-L'outil scatter matrix



Dans une matrice de diffusion, chaque élément de matrice E_{ij} est un diagramme de dispersion des colonnes i et j , où les valeurs de la i -ème colonne sont affichées sur l'axe x et les valeurs de la j -ème colonne sur l'axe y tandis que les coordonnées sont affichées en alternance sur tous les côtés de l'intrigue.



3-Conclusion

La visualisation des données est définie comme l'exploration visuelle et interactive de données de toutes volumétries, natures (structurées ou non structurées) et origines, et leur représentation graphique. et aident à voir des choses qui n'étaient pas évidentes auparavant. Même quand le volume des données est très important comme dans notre exemple où on trouve 195 données déferents de 31 personnes ou des tendances peuvent être perçues de façon rapide et simple.

